Applied Linguistics: 31/4: 487–512

doi:10.1093/applin/amp058 Advance Access published on 12 January 2010

An Academic Formulas List: New Methods in Phraseology Research

RITA SIMPSON-VLACH and NICK C. ELLIS

University of Michigan

This research creates an empirically derived, pedagogically useful list of formulaic sequences for academic speech and writing, comparable with the Academic Word List (Coxhead 2000), called the Academic Formulas List (AFL). The AFL includes formulaic sequences identified as (i) frequent recurrent patterns in corpora of written and spoken language, which (ii) occur significantly more often in academic than in non-academic discourse, and (iii) inhabit a wide range of academic genres. It separately lists formulas that are common in academic spoken *and* academic written language, as well as those that are special to academic written language alone and academic spoken language alone. The AFL further prioritizes these formulas using an empirically derived measure of utility that is educationally and psychologically valid and operationalizable with corpus linguistic metrics. The formulas are classified according to their predominant pragmatic function for descriptive analysis and in order to marshal the AFL for inclusion in English for Academic Purposes instruction.

AN ACADEMIC FORMULAS LIST

The aim of this research is to create an empirically derived and pedagogically useful list of formulaic sequences for academic speech and writing, comparable with the Academic Word List (hereafter AWL; Coxhead 2000). It is motivated by current developments in language education, corpus linguistics, cognitive science, second language acquisition (SLA), and English for academic purposes (EAP). Research and practice in SLA demonstrates that academic study puts substantial demands upon students because the language necessary for proficiency in academic contexts is quite different from that required for basic interpersonal communicative skills. Recent research in corpus linguistics analyzing written and spoken academic discourse has established that highly frequent recurrent sequences of words, variously called lexical bundles, chunks, multiword expressions (*inter alia*) are not only salient but also functionally significant. Cognitive science demonstrates that knowledge of these formulas is crucial for fluent processing. And finally, current trends in SLA and EAP demand ecologically valid instruction that identifies and prioritizes the most important formulas in different genres.

The AFL includes formulaic sequences, identifiable as frequent recurrent patterns in written and spoken corpora that are significantly more common in academic discourse than in non-academic discourse and which occupy a

range of academic genres. It separately lists formulas that occur frequently in both academic spoken and academic written language, as well as those that are more common in either written or spoken genres. A major novel development this research brings to the arena is a ranking of the formulas in these lists according to an empirically derived psychologically valid measure of utility, called 'formula teaching worth' (FTW). Finally, the AFL presents a classification of these formulas by pragma-linguistic function, with the aim of facilitating their inclusion in EAP curricula.

BACKGROUND

Functional, *cognitive linguistic* and *usage-based theories* of language suggest that the basic units of language representation are *constructions*—form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind (Langacker 1987; Tomasello 1998, 2003; Barlow and Kemmer 2000; Croft and Cruise 2004; Goldberg 2006; Robinson and Ellis 2008;). Constructions are associated with particular semantic, pragmatic, and discourse functions, and are acquired through engaging in meaningful communication. Constructions form a structured inventory of a speaker's knowledge of the conventions of their language, as independently represented units in a speaker's mind. Native-like selection and fluency relies on knowledge and automatized processing of these forms (Pawley and Syder 1983; Ellis 2009).

Corpus Linguistics confirms the recurrent nature of these formulas (Hunston and Francis 1996; McEnery and Wilson 1996; Biber *et al*. 1998). Large stretches of language are adequately described as collocational streams where patterns flow into each other. Sinclair (1991, 2004) summarizes this in his *'idiom principle:'* 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments.' (1991: 110). Rather than being a minor feature, compared with grammar, Sinclair suggests that for normal texts, the first mode of analysis to be applied is the idiom principle, as most text is interpretable by this principle. Comparisons of written and spoken corpora demonstrate that more collocations are found in spoken language (Brazil 1995; Biber *et al*. 1999; Leech 2000). Speech is constructed in real time and this imposes greater working memory demands than writing, hence the greater the need to rely on formulas: it is easier to retrieve something from long-term memory than to construct it anew (Kuiper 1996; Bresnan 1999).

Many formulaic constructions are non-compositional or idiomatic, like 'once upon a time', or 'on the other hand', with little scope for substitution ('twice upon a time', 'on the other foot') (Simpson and Mendis 2003). Even those that appear to be more openly constructed may nevertheless be preferred over alternatives (in speech, 'in other words' 'to say it differently',

'in paraphrase', '*id est*') with the demands of native-like selection entailing that every utterance be chosen from a wide range of possible expressions, to be appropriate for that idea, for that speaker, for that genre, and for that time. Natives and experts in particular genres learn these sequential patterns through repeated usage (Pawley and Syder 1983; Ellis 1996, 2009; Wray 1999, 2002). Psycholinguistic analyses demonstrate that they process collocations and formulas with greater facility than 'equivalent' more open constructions (Bybee and Hopper 2001; Ellis 2002a, 2002b; Jurafsky 2002; Bod *et al*. 2003; Schmitt 2004; Ellis *et al*. 2008, 2009). For example, in speech production, 'items that are used together, fuse together' (Bybee 2003: 112): words that are commonly uttered in sequence become pronounced as a whole that is shortened and assimilated ('give + me' → 'gimme'; 'I + am + going + to' → 'I'm gonna', etc.). The phenomenon is graded—the degree of reduction is a function of the frequency of the target word and the conditional probability of the target given the surrounding words (Bybee and Hopper 2001; Jurafsky *et al*. 2001).

EAP research (e.g. Swales 1990; Flowerdew and Peacock 2001; Hyland 2004, 2008; Biber and Barbieri 2006) focuses on determining the functional patterns and constructions of different academic genres. These analyses have increasingly come to be based on corpora representative of different academic fields and registers, such as the Michigan Corpus of Academic Spoken English (Simpson *et al*. 2002), with qualitative investigation of patterns, at times supported by computer software for analysis of concordances and collocations. But these studies need to be buttressed with quantitative information too, as in the case of vocabulary where there have been longstanding attempts to identify the more frequent words specific to academic discourse and to determine their frequency profile, harking back, for example, to the University Word List (West 1953). The logic for instruction and testing is simple—the more frequent items have the highest utility and should therefore be taught and tested earlier (Nation 2001).

The most significant recent developments in this direction have been those of Coxhead (2000). Her development of the AWL has had a significant impact on EAP teaching and testing because it collects words that have high currency in academic discourse by applying specific criteria of frequency and range of distribution in a 3.5-million-word corpus of academic writing representing a broad spectrum of disciplines. Because academic study puts unique demands on language learners, the creation of the AWL as a teaching resource filled a substantial gap in language education by providing a corpus-based list of lexical items targeted specifically for academic purposes.

Can the same principles of academic vocabulary analysis be applied to other lexical units characterizing academic discourse? Can the theoretical research on formulaic language, reviewed above, which demonstrates that contiguous multiword phrases are important units of language, be likewise transformed

into practical pedagogical uses (Nattinger and DeCarrico 1992; Lewis 1993; Wray 2000; Schmitt 2004)? Is an AFL equally viable?

A crucial factor in achieving this goal lies in the principles for identifying and classifying such units. The *lexical bundle* approach of Biber and colleagues (1998, 2004), based solely on frequency, has the advantage of being methodologically straightforward, but results in long lists of recurrent word sequences that collapse distinctions that intuition would deem relevant. For example, few would argue with the intuitive claim that sequences such as 'on the other hand' and 'at the same time' are more psycholinguistically salient than sequences such as 'to do with the', or 'I think it was', even though their frequency profiles may put them on equivalent lists. Selection criteria that allow for intuitive weeding of purely frequency-based lists, as used by Simpson (2004) in a study of formulaic expressions in academic speech, yield much shorter lists of expressions that may appeal to intuitive sensibilities, but they are methodologically tricky and open to claims of subjectivity.

In this paper, we present a method for deriving a list of formulaic expressions that uses an innovative combination of quantitative and qualitative criteria, corpus statistics and linguistic analyses, psycholinguistic processing metrics, and instructor insights. Long lists of highly frequent expressions are of minimal use to instructors who must make decisions about what content to draw students' attention to for maximum benefit within limited classroom time. The fact that a formula is above a certain frequency threshold and distributional range does not necessarily imply either psycholinguistic salience or pedagogical relevance; common sequences of common words, such as 'and of the,' are expected to occur frequently. Psycholinguistically salient sequences, on the other hand, like 'on the other hand', cohere much more than would be expected by chance; they are 'glued together' and thus measures of association, rather than raw frequency, are likely more relevant. Our primary aim in this research is to create a pedagogically useful list of formulaic sequences for academic speech and writing. A secondary aim, however, is to discuss the statistical measures beyond frequency counts available for ranking formulaic sequences extracted from a corpus. The departure point for our research was dissatisfaction with a strictly frequency-based rank ordering of multiword phrases on the one hand, and a frequency plus intuition-based ordering on the other hand, coupled with a need for relatively contained, manageable sets of multiword expressions for use in classroom applications and teaching materials development. We used frequency as a starting point, but our approach is substantially more robust than the previous corpus-based methods for classifying multiword formulas; it encompasses a statistical measure of cohesiveness—mutual information (MI)—that has heretofore not been used in related research, in conjunction with validation and prioritization studies designed to provide insights into which formulas are perceived to be the important ones for teaching.

METHODS

The corpora

Target corpora

The target corpora of academic discourse included 2.1 million words each of academic speech and academic writing. The academic speech corpus was comprised of MICASE (1.7 million words) (Simpson *et al*. 2002) plus BNC files of academic speech (431,000 words) (British National Corpus 2006). The academic writing corpus consisted of Hyland's (2004) research article corpus (1.2 million words), plus selected BNC files (931,000 words) sampled across academic disciplines using Lee's (2001) genre categories for the BNC.[1] The speech corpus was broken down into five subcorpora and the writing corpus into four subcorpora by academic discipline, as shown in Table 1.

Comparison corpora

For comparative purposes, two additional corpora were used. For non-academic speech, we used the Switchboard (2006) corpus (2.9 million words), and for non-academic writing we used the FLOB and Frown corpora (1.9 million words) which were gathered in 1991 to reflect British and American English over 15 genres and to parallel the original LOB and Brown collections (ICAME 2006). FLOB and Frown were favored over their predecessors because the age of the texts is closer to the target corpus texts. The Switchboard corpus was chosen because it contains unscripted casual telephone conversations, and thus lies near the opposite end of the style spectrum from academic speech.[2]

Formula identification and MI

The first decision was what length of formulas we would include in the data. It is well known that 2-word phrases (bi-grams) are highly frequent and

Table 1: Word counts by discipline for the Academic subcorpora

Academic speech		Academic writing	
Discipline	Word count	Discipline	Word count
Humanities and Arts	559,912	Humanities and Arts	360,520
Social Sciences	710,007	Social Sciences	893,925
Biological Sciences	357,884	Natural Sciences/Medicine	513,586
Physical Sciences	363,203	Technology and Engineering	349,838
Non-departmental/other	159,592		
Total	2,153,770	Total	2,117,869

include many phrases that are subsumed in 3- or 4-word phrases; so we excluded 2-word sequences, to keep the data set to a more manageable size. Although recurrent 5-word sequences are comparatively rare, we decided to include them for the sake of thoroughness, thus including strings of 3, 4, and 5 words into the data set. The next decision was what frequency level to use as a cutoff. Previous research uses cutoff ranges between 10 and 40 instances per million words. Since our research goals included using other statistical measures to cull and rank the formulas, we wanted a less restricted data set to start with, and so opted for the lowest frequency range used in previous research, namely 10 per million (Biber *et al*. 1999).

We began by extracting all 3-, 4-, and 5-grams occurring at least 10 times per million from the two target and two comparison corpora, using the program *Collocate* (Barlow 2004). These four data sets naturally included a great deal of overlap, but also substantial numbers of phrases unique to each corpus. The next step then was to collapse the overlapping data and collect frequency counts for each phrase appearing in any one of those four corpora (at the threshold level of 10 per million) for all the other corpora, for comparison purposes. The total number of formulas in this list was approximately 14,000.

From this master list, we wanted to determine which formulas were more frequent in the academic corpora than in their non-academic counterparts, because our goal was to identify those formulas that are characteristic of academic discourse in particular, in contrast to high-frequency expressions occurring in any genre. This is an important step that warrants additional justification. Just as the AWL omitted words that were in the most frequent 2,000 words of English, we needed a way to sift out the most frequent formulas occurring in both academic and non-academic genres. To accomplish this, we used the log-likelihood (LL) statistic to compare the frequencies of the phrases across the academic and non-academic corpora. The LL ratio is useful for comparing the relative frequency of words or phrases across registers and determining whether the frequency of an item is statistically higher in one corpus or subcorpus than another (Oakes 1998; Jurafsky and Martin 2000; Rayson and Garside 2000). Those expressions found to occur statistically more frequently in academic discourse, using the LL statistic with a significance level of $p = 0.01$, comprise the basis for the academic formulas list (AFL). We separately compared academic vs. non-academic speech, resulting in over 2,000 items, and academic vs. non-academic writing, resulting in just under 2,000 items. The overlapping items from these two lists were identified as the core formulas that appear frequently in both academic speech and writing.

Once these lists were obtained, cutoff values for distributional range across the academic subdivisions of the corpora had to be established. The subcorpora for academic speech were (Table 1): Humanities and Arts, Social Sciences, Biological and Health Sciences, Physical Sciences and Engineering, and Other/non-disciplinary. For academic writing, the subcorpora were: Humanities and Arts, Social Sciences, Natural Sciences and Medicine, and

Technology and Engineering. The cutoff values we used were as follows: Expressions occurring primarily in speech had to occur at the 10 tokens per million level or above in *four out of five* of the academic divisions, resulting in a Spoken AFL of 979 items; expressions occurring primarily in writing had to occur at least 10 times per million words in *three out of four* academic divisions, resulting in a Written AFL of 712 items; and expressions occurring in both speech and writing had to occur at a level of 10 per million in at least *six out of all nine* subcorpora, resulting in a Core AFL of 207 items.[3] These range thresholds ensure that the AFL formulas are found across the breadth of academic spoken or written language and are thus relevant to general EAP, rather than to particular disciplines. Furthermore, the range ensures that the formulas on the list are not attributable to the idiosyncrasies of particular speakers or speech events.

Another important statistic we calculated for each of the strings was the MI score. MI is a statistical measure commonly used in the field of information science designed to assess the degree to which the words in a phrase occur together more frequently than would be expected by chance (Oakes 1998; Manning and Schuetze 1999). A higher MI score means a stronger association between the words, while a lower score indicates that their co-occurrence is more likely due to chance. MI is a scale, not a test of significance, so there is no minimum threshold value; the value of MI scores lies in the comparative information they provide. The question we then posed is: To what extent are these corpus metrics of frequency and MI useful for ranking the formulas on a list?

High frequency *n*-grams occur often. But this does not imply that they have clearly identifiable or distinctive functions or meanings; many of them occur simply by dint of the high frequency of their component words, often grammatical functors. In addition, relying solely on frequency means that some distinctively useful but lower frequency phrases whose component words are highly unlikely to occur together by chance will not make it to the top of the frequency-ordered list. So frequency alone is not a sufficient metric.

High MI *n*-grams are those with much greater coherence than is expected by chance, and this tends to correspond with distinctive function or meaning. But this measure tends, in contrast to frequency, to identify rare phrases comprised of rare constituent words, such as many subject-specific phrases. So nor is MI alone a perfect metric for extracting phrases that are highly noteworthy for teachers, since it privileges low-frequency items. Tables 2 and 3 present a simple re-ordering by frequency and MI of the top 10 and bottom 10 phrases of the approximately 2,000 original Academic speech and original Academic writing items to illustrate these points.

For the speech data in Table 2, we see that frequency prioritizes such phrases as 'and this is' and 'this is the' which seem neither terribly functional nor pedagogically compelling, while it satisfactorily relegates to the bottom the phrases 'cuz if you', and 'um and this'. Instructors might, however, be interested in

Table 2: The top 10 and bottom 10 phrases of the original Academic speech items prioritized by frequency and by MI

Top 10 by frequency	**Top 10 by MI**
this is the	blah blah blah
be able to	trying to figure out
and this is	do you want me to
you know what	for those of you who
you have a	we're gonna talk about
you can see	talk a little bit
look at the	does that make sense
you need to	thank you very much
so this is	the university of Michigan
you want to	you know what i mean
Bottom 10 by frequency	**Bottom 10 by MI**
if you haven't	okay and the
so what we're	is like the
as well but	so in the
cuz if you	and so the
right okay and	the um the
um and this	is what the
think about how	this in the
we're interested in	that it's the
will give you	is it the
we can we	of of of

other low frequency neighbors such as 'we're interested in', and 'think about how'. MI, on the other hand, privileges functional formulas such as 'does that make sense' and 'you know what I mean', though 'blah blah blah' and 'the University of Michigan' are high on the list too. The low priority items by MI such as 'the um the' and 'okay and the' do indeed seem worthy of relegation. For the written data in Table 3, frequency highlights such strings as 'on the other hand' and 'it is possible' (we think appropriately), alongside 'it has been' and 'but it is' (we think inappropriately), and pushes 'by the use' and 'of the relevant' to the bottom (appropriately), alongside 'it is obvious that' and 'in the present study' (inappropriately). MI, in contrast, prioritizes such items as 'due to the fact that' and 'there are a number of' (appropriately; indeed all of the top ten seem reasonable), and it (appropriately) relegates generally non-functional phrases such as 'to be of', 'as to the', 'of each of', etc. These tables represent just a glimpse of what is revealed by the comparison of a given list of formulas ordered by these two measures. Our intuitive impressions of the prioritizations produced by these measures on their own, as illustrated here, thus led us to

Table 3: The top 10 and bottom 10 phrases of the original Academic writing items prioritized by frequency and by MI

Top 10 by frequency	**Top 10 by MI**
on the other	due to the fact that
in the first	it should be noted
the other hand	on the other hand the
on the other hand	it is not possible to
in the united	there are a number of
but it is	in such a way that
can be seen	a wide range of
it has been	take into account the
is likely to	on the other hand
it is possible	as can be seen
Bottom 10 by frequency	**Bottom 10 by MI**
is sufficient to	to the case
weight of the	of each of
of the relevant	with which the
by the use of	as in the
the assessment of	it is of
by the use	is that of
of the potential	to that of
it is obvious that	as to the
in the present study	to be of
is obvious that	that as the

favor MI over frequency. Ideally, though, we wanted to combine the information provided by *both* metrics to better approximate our intuitions and those of instructors, and thus to rank the academic formulas for use in pedagogical applications.

Our efforts to achieve this synthesis were part of a large validation study which triangulated corpus linguistic measures, educator insights, and psycholinguistic processing measures. A full description of these investigations is available in Ellis *et al*. (2008). Because these details are available elsewhere, and because the primary aim of the present paper is to present the AFL items and their functional categorizations, we simply summarize the relevant parts of the procedures here.

Determining a composite metric to index FTW

We selected a subset of 108 of these academic formulas, 54 from the spoken and 54 from the written list. These were chosen by stratified random sampling

to represent three levels on each of three factors: *n*-gram length (3,4,5), frequency band (High, Medium, and Low; means 43.6, 15.0, and 10.9 per million, respectively), and MI band (High, Medium, and Low; means 11.0, 6.7, and 3.3, respectively). There were two exemplars in each of these cells.

We then asked twenty experienced EAP instructors and language testers at the English Language Institute of the University of Michigan to rate these formulas, given in a random order of presentation, for one of three judgements using a scale of 1 (disagree) to 5 (agree):

A. whether or not they thought the phrase constituted 'a formulaic expression, or fixed phrase, or chunk'. There were six raters with an inter-rater $\alpha = 0.77$.
B. whether or not they thought the phrase has 'a cohesive meaning or function, as a phrase'. There were eight raters with an inter-rater $\alpha = 0.67$.
C. whether or not they thought the phrase was 'worth teaching, as a bona fide phrase or expression'. There were six raters with an inter-rater $\alpha = 0.83$.

Formulas which scored high on one of these measures tended to score high on another: r AB = 0.80, $p < 0.01$; r AC = 0.67, $p < 0.01$; r BC = 0.80, $p < 0.01$). The high alphas of the ratings on these dimensions and their high inter-correlation reassured us of the reliability and validity of these instructor insights. We then investigated which of frequency or MI better predicted these instructor insights. Correlation analysis suggested that while both of these dimensions contributed to instructors valuing the formula, it was MI which more strongly influenced their prioritization: r frequency/A = 0.22, $p < 0.05$; r frequency/B = 0.25, $p < 0.05$; r frequency/C = 0.26, $p < 0.01$; r MI/A = 0.43, $p < 0.01$; r MI/B = 0.51, $p < 0.01$; r MI/C = 0.54, $p < 0.01$. A multiple regression analysis predicting instructor insights regarding whether an *n*-gram was worth teaching as a bona fide phrase or expression from the corpus metrics gave a standardized solution whereby teaching worth = β 0.56 MI + β 0.31 frequency. That is to say, when instructors judge *n*-grams in terms of whether they are worth teaching, considering both frequency and MI factor into their judgements, it is the MI of the string—the degree to which the words are bound together—that is the major determinant.

These beta coefficients, derived from the 108 formula subset for which we had obtained instructor ratings, could then be used over the population of academic formulas which they represented to estimate from the two corpus statistics available for all formulas—the combined measures of MI and frequency—a FTW score that is a prediction of how instructors would judge their teaching worth. This score, like the MI statistic, does not provide a threshold cutoff score, but enables a reliable and valid rank ordering of the formulas, which in turn provides instructors and materials developers with a basis for prioritizing formulaic expressions for instructional uses. The FTW score, with its use of both frequency rank and MI score is thus a methodologically

innovative approach to the classification of academic formulas, as it allows for a prioritization based on statistical and psycholinguistic measures, which a purely frequency-based ordering does not.

RESULTS: THE AFL AND FUNCTIONAL CATEGORIZATION

In the appendix (see supplementary material available at Applied Linguistics online), we present the AFL grouped into the three sublists— the Core AFL in its entirety, and the first 200 formulas of the Spoken AFL, and the Written AFL. Since all three lists are sorted by the two-factor FTW score, providing the top 200 formulas for the two longer AFL components effectively distills them into the most relevant formulas.

Scrutiny of the lists also shows substantial overlap among some of the entries. Thus, for example, in Appendix 1, the Core AFL listing includes the *n*-grams *from the point of view*, *the point of view of*, *point of view*, *point of view of*, *the point of view*, etc. Since this degree of redundancy is not especially useful, and moreover takes up extra space, in our functional categorizations we collapsed incidences like these together into their common schematic core—in this case, (*from*) (*the*) *point of view* (*of*). We retained the original formulas in the Appendix 1 tables, but only collapsed them in Table 4, the functional categorization. We acknowledge that in so doing we have sacrificed some detail as to the specific configurations and functions of component phrases; however, the differences in pragmatic function of these formula variations are generally minor and the detail lost can easily be retrieved by looking at the fuller lists in the appendix.

The final stage of the analysis involved grouping the formulas into categories according to their primary discourse-pragmatic functions. For purposes of expediency as well as the anticipated pedagogical applications, we again included only those formulas from the Core AFL list and the top 200 from the Written AFL and the Spoken AFL lists. These functional categories—determined after examining the phrases in context using a concordance program—are not meant to be taken as definitive and exclusive, since many of the formulas have multiple functions, but rather as indications of the most salient function the phrases fulfill in academic contexts. In the following section, we present an overview of the functional analysis, providing examples to illustrate some of the more important functions in context.

Rationale and overview of the functional categories

The purpose of the following classification is primarily pedagogical. An ordered list of formulas sorted according to major discourse-pragmatic functions allows teachers to focus on functional language areas which, ideally, will dovetail with functional categories already used in EAP curricula. The creation of a functional taxonomy for formulaic sequences is an inherently problematic endeavor, as Wray and Perkins (2000: 8) point out, arguing that typologies

Table 4: The AFL categorized by function

Group A. Referential expressions

(1) Specification of attributes

(a) Intangible framing attributes

Core AFL (written & spoken)

[a/the] form of	(in) such a (way)	the distribution of	the problem of
(as) a function (of)	(in) terms of (the)	the existence of	the process of
based on [a/the]	in which the	(the) extent to which	the question of
focus on the	is based on (the)	(the) fact that (the)	the role of
form of the	nature of the	the idea that	the structure of
(from) (the) point of	of the fact	the issue of	the study of
view (of)	(on) the basis (of)	the meaning of	(the) way(s) in (which)
in relation to	the ability to	the nature of (the)	the way that
in response to	the concept of	the notion of	the work of
(in) the case (of)	the context of	the order of	the use of
in the context (of)	the definition of	the presence of (a)	with respect to (the)
in the sense (that)	the development of		

Primarily spoken

it in terms of	the idea of	the kind of	this kind of

Primarily written

an attempt to	in accordance with (the)	in the course of	on the basis of the
[are/was] based on	(in) such a way that	in the form of	on the part of
by virtue of	in terms of a	in this case the	to the fact that
degree to which	in the absence of	insight into the	with regard to
depend([ing/s]) on the			

(b) Tangible framing attributes

Core AFL (written & spoken)

(as) part of [a/the]	the change in	(the) part(s) of the	(the) size of (the)
the amount of	the frequency of	the rate of	(the) value of (the)
the area of	the level of	the sum of	

Written AFL

an increase in the	High levels of	over a period of	

(c) Quantity specification

Core AFL (written & spoken)

a list of	[a/large/the] number of	both of these	of the second
a series of	And the second	each of [the/these]	the first is
a set of		of [the/these] two	there are three

Primarily spoken

all sorts of			

Primarily written

a high degree	little or no	in some cases	there are no
a large number (of)	in a number of	(the) total number (of)	there are several
(a) small number (of)	in both cases	(there) are a number (of)	two types of
(a) wide range (of)	in most cases		

(continued)

Table 4: Continued

(2) Identification and focus

Core AFL (written & spoken)

a variety of	is for the	it is not	that this is
[an/the] example of (a)	is not [a/the]	means that the	that we are
as an example	is that [it/the/there]	referred to as	there is [a/an/no]
different types of	is the case	such as the	this is [a/an/not]
here is that	is to be	that in [a/the]	this type of
if this is	it can be	that is the	this would be
	it does not	that there [are/is (a)]	which is [not/the]

Primarily spoken

[has/have] to do with	how many of you	so this is	this is the
it's gonna be	nothing to do with	the best way to	this is this is
and this is	one of these	there was a	those of you who
for those of you (who)			

Primarily written

(as) can be seen (in)	it has been	that there is no	this does not
does not have	none of these	there has been	this means that
has also been	that it is not	they [did/do] not	which can be
his or her			

(3) Contrast and comparison

Core AFL (written & spoken)

and the same	different from the	is much more	(the) difference between (the)
as opposed to	exactly the same	related to the	the relationship between
associated with the	have the same	the same as	
between the two	[in/of/with] the same		

Primarily spoken

(nothing) to do with (the)	the same thing	to each other	

Primarily written

be related to the	(on) the other (hand)	the difference between	(the) same way as
is more likely	(the)	the	to distinguish between
	similar to those		

(4) Deictics and locatives

Core AFL (written & spoken)

a and b	the real world	of the system	

Primarily spoken

(at) the end (of) (the)	(at) (the) University of	in Ann Arbor	piece of paper
at this point	Michigan		

Primarily written

at the time of	at this stage	b and c	the United Kingdom

(5) Vagueness markers

Core AFL (written & spoken)

and so on

Primarily spoken

and so forth	and so on and so	blah blah blah	

(continued)

Table 4: Continued

Group B. Stance expressions

(1) Hedges

Core AFL (written & spoken)

(more) likely to (be)	[it/there] may be	may not be	to some extent

Primarily spoken

a kind of	it could be	it might be	might be able (to)
a little bit about	it looks like	little bit about	you might want to
in a sense			

Primarily written

appear(s) to be	at least in	is likely to (be)	it is likely that
are likely to	does not appear	it appears that	less likely to
as a whole			

(2) Epistemic stance

Core AFL (written & spoken)

according to the	assume that the	to show that	we can see
be the case	out that the		

Primarily spoken

[and/as] you can (see)	how do we	trying to figure (out)	what do you mean
do you know what	how do you know	to figure out (what)	what does that mean
(does) that make sense	I think this is	you think about it	(you) know what I
		okay I don't know	(mean)

Primarily written

assumed to be	be seen as	be considered as	is determined by
be argued that	been shown to	have shown that	we assume that
be explained by	can be considered	if they are	we have seen
be regarded as			

(3) Obligation and directive

Primarily spoken

do you want (me) (to)	I want you to	tell me what	you don't need to
doesn't have to be	it has to be	(to) make sure (that)	you need to (do)
don't worry about	keep in mind	we have to	you want me to
has to be	take a look (at)	we need to	you want to

Primarily written

(it should) be noted	need not be	should also be	take into account (the)
(that)	needs to be	should not be	to ensure that (the)

(4) Expressions of ability and possibility

Core AFL (written & spoken)

can be used (to)	to use the

Primarily spoken

(gonna) be able (to)	that you can	(you) can look at	you could you could
so you can (see)	to think about	you can see ([that/the])	you're trying to

Primarily written

allows us to	be used as a	can easily be	it is possible ([that/to])
are able to	be used to	can be found (in)	most likely to
be achieved by	can also be	could be used	their ability to
[be/been/was] carried out	can be achieved	has been used	to carry out
carried out [by/in]	can be expressed	(it) is not possible (to)	

(continued)

Table 4: Continued

(5) Evaluation

Core AFL (written & spoken)
the importance of

Primarily Spoken
it doesn't matter

Primarily written

important role in	it is important (to)	it is necessary (to)	(it) is clear (that)
is consistent with	it is impossible to	it is obvious that	the most important
it is difficult	it is interesting to	it is worth	

(6) Intention/volition, prediction

Primarily spoken

I just wanted to	if you wanna	if you were (to)	I'm not gonna
I wanted to	if you want(ed) (to)	I'm gonna go	let me just
			um let me

Primarily written

to do so	we do not

Group C: Discourse organizing functions

(1) Metadiscourse and textual reference

Primarily spoken

come back to	I'm talking about	we talk(ed) about	We've talked about
go back to the	talk a little bit	we were talking (about)	what I'm saying
gonna talk about	talk(ing) about the	We'll talk about	what I'm talking about
I was gonna say	to talk about	We're gonna talk (about)	what you're saying
(I) was talking about	wanna talk about	We're talking about	You're talking about
I'll talk about			

Primarily written

as shown in	in the next section	(in) this paper (we)	shown in table
at the outset	in the present study	shown in figure	the next section
in table 1	in this article		

(2) Topic introduction and focus

Core AFL (written & spoken)

For example [if/in/the]	what are the

Primarily spoken

a look at	if you've got	wanna look at	when you look at
first of all	let's look at	we look(ed) at	you have a
I have a question	look at [it/the/this]	we're looking at	you look at (the)
I'll show you	looking at the	what I mean	you're looking at
if you have (a)	to look at (the)	what I want to	you've got a
if you look (at) (the)			

(3) Topic elaboration

(a) non-causal

Core AFL (written & spoken)
But this is

Primarily spoken

any questions about	I mean if (you)	see what I'm saying	what happens is
came up with	(it) turns out (that)	so if you	you know what I'm
come up with (a)			

(continued)

Table 4: Continued

Primarily written			
are as follows	in more detail	see for example	such as those
factors such as			
(b) Topic elaboration: cause and effect			
Core AFL (written & spoken)			
[a/the] result of	due to the	so that the	the reason for
(as) a result (of)	in order to	the effect(s) of	whether or not (the)
because it is			
Primarily spoken			
End up with	in order to get	the reason why	
Primarily written			
as a consequence	for the purposes of	give rise to	it follows that
as a result of the	for this purpose	is affected by	to determine whether
due to the fact (that)	for this reason		
(4) Discourse markers			
Core AFL (written & spoken)			
and in the	as well as	at the same (time)	(in) other words (the)
Primarily spoken			
and if you	but if you	no no no (no)	oh my god
and then you	by the way	thank you very (much)	yes yes yes
Primarily written			
even though the	in conjunction with		

The table includes all 207 formulas of the Core List, the top 200 items of the Written AFL and the top 200 items of the Spoken AFL lists.

such as those offered by Nattinger and DeCarrico (1992), among others, suffer from a proliferation of types and subtypes. This proliferation of categories does indeed make it difficult to distill the data into a compact functional model applicable across corpora and domains of use. In spite of these difficulties, however, we maintain that for pedagogical purposes, a functional taxonomy, however multilayered or imprecise because of overlapping functions and multifunctional phrases, is nevertheless crucial to enhancing the usefulness of the AFL for teachers. As for pedagogical applications, this functional categorization of the AFL is intended primarily as a resource for developing teaching materials based on further contextual research around the items rather than a resource for teaching itself. Due to space constraints, we cannot present specific teaching suggestions here, but do reiterate that the formula in context is what is pedagogically relevant. The functional categorization of the AFL is an important resource, but nevertheless only a starting point.

Previous researchers have in fact already paved the way in this area; in particular, we credit the work of Biber *et al*. (2004) in this aspect of our study. The current classification scheme is an adaptation of the functional taxonomy outlined in their article, but with some important extensions and

modifications. As in their study, we grouped the formulas into three primary functional groups: referential expressions, stance expressions, and discourse organizers.

Several functional categories in our classification scheme, however, are not in the Biber *et al*. taxonomy, and these should be mentioned here. Within the referential expressions group, we have added one category—namely, that of contrast and comparison. This is a common functional category in EAP curricula, and with over 20 formulas it represents an important functional group of the AFL. For the category of stance expressions, a number of formulas represent two essential categories not explicitly named by Biber *et al*.: These are hedges and boosters, and evaluation. In addition, we have collapsed two of their categories (desire and intention/prediction) into one, called volition/intention, since the AFL formulas in the two categories did not seem distinct enough in their discourse functions to warrant splitting them. Finally, the discourse organizers group is substantially expanded and modified from the Biber *et al*. grouping, with three important additional subcategories: metadiscourse and textual reference, cause and effect expressions, and discourse markers. Our functional classification is thus considerably more extensive than Biber *et al*.'s; we suspect that this may be due primarily to the fact that there are close to 500 formulas in this portion of the AFL, compared with fewer than 150 phrases included in their list of the most common lexical bundles. Finally, we reiterate that even though some of the formulas are multifunctional, we have nevertheless tried to align all of them with their most probable or common function.

Description and examples of the functional categories

The following section outlines the pragmatic functional taxonomy. Numbers in brackets refer to the total number of formulas in that category from the combined Core AFL and top 200 each from the Written AFL and Spoken AFL.

Group A: Referential expressions

The largest of the three major functional groupings, the referential expressions category encompasses five subcategories: specification of attributes, identification and focus, contrast and comparison, deictics and locatives, and vagueness markers.

(1) Specification of attributes

(a) Intangible framing attributes [66]. The largest pragmatic subcategory for all AFL phrases is the specification of attributes—intangible framing devices. The majority of these phrases appear on the Core AFL list, indicating that these are clearly important academic phrases across both spoken and written genres. This category includes phrases that frame both concrete entities (as in A.1) and abstract concepts or categories (as in A.2).

(A.1) . . . *based on the* total volume passing through each cost center
(A.2) so even with *the notion of* eminent domain and fair market value . . .

There are close to 70 formulas in this category, and roughly half are composed of the structure *'a/the N of'*, sometimes with a preceding preposition, as in *as a function of, on the basis of*, and *in the context of*. Most of these formulas frame an attribute of a following noun phrase, but some frame an entire clause (A.3), or function as a bridge between a preceding verb and a following clause (A.4).

(A.3) But another clear example of *the way in which* domestic and foreign policy overlaps is of course in economic affairs.
(A.4) human psychology has evolved *in such a way*, as to allow us to make those kinds of judgements that would normally be reliable.

(b) Tangible framing attributes [14]. The second subcategory of attribute specifiers is that of tangible framing attributes such as *the amount of, the size of, the value of*, which refer to physical or measurable attributes of the following noun.

(A.5) this is uh, what she found in terms of *the level of* shade and yield of coffee . . .

(c) Quantity specification [26]. The final subcategory of attribute specifiers is closely related to the category of tangible framing attributes, and includes primarily cataphoric expressions enumerating or specifying amounts of a following noun phrase, as in *a list of, there are three, little or no, all sorts of*. Some of the quantity specifiers, however—for example, *both of these, of these two*—are anaphoric, referring to a prior noun phrase (e.g. A.7).

(A.6) From an instrumental viewpoint, *there are three* explanations worth considering.
(A.7) It is the combination *of these two* that results in higher profits to the EDLP store.

(2) Identification and focus [53]. The second most common functional category, with 53 formulas, is the subcategory of identification and focus, which includes typical expository phrases such as *as an example, such as the, referred to as*, and *means that the*, and also a number of stripped-down sentence or clause stems with a copula, auxiliary verb, or modal construction, such as *it is not, so this is, this would be*. It is not surprising that this functional category figures prominently in academic discourse, since exemplification and identification are basic pragmatic functions in both academic speech and writing. In fact, these phrases often occur in clusters, as in example A.9.

(A.8) So many religions, *such as the* religion of Ancient Egypt, for instance . . .
(A.9) so *this would be an example of* peramorphosis.

(3) Contrast and comparison [23]. Many of the contrast and comparison phrases included explicit markers of comparison such as *same*, *different*, *or similar*. As mentioned earlier, this category is not included in Biber *et al.*, but constitutes an important language function for EAP teaching purposes.

(A.10) that's probably a prefix code *as opposed to* a suffix code.

(4) Deictics and locatives [12]. The deictic and locative expressions are a small but important functional category, referring to physical locations in the environment (e.g. *the real world)* or to temporal or spatial reference points in the discourse (e.g. *a and b*, *at this point)* These formulas obviously reflect the provenance of the corpus, so *the University of Michigan*, *Ann Arbor,* and *the United Kingdom* all appear on this list because of the inclusion of both MICASE and BNC texts.

(5) Vagueness markers [4]. There are only four phrases included in the AFL that are classified as vagueness markers, making it the smallest functional category. Furthermore, three of these phrases are limited to the Spoken AFL; only the phrase *and so on* appears in the Core AFL. Nevertheless, the frequency rates and FTW scores show that these phrases are important; making vague references with these particular extenders is a common discourse function in academic speech. Interestingly, Biber *et al*. (2004) also only list three phrases in this category (which they call imprecision bundles), yet claim that it is a major subcategory of referential bundles; perhaps this claim is also based on frequencies. Note that the three phrases they list in this category (*or something like that*, *and stuff like that*, *and things like that*), do not appear in the AFL, because although they may indeed be frequent in academic speech, they were not sufficiently *more* frequent in academic speech as compared with non-academic speech to make the cut for the AFL.

Group B: Stance expressions

Stance formulas include six functional subcategories, two of which—hedges and evaluative formulas—are additions to the Biber *et al*. taxonomy.

(1) Hedges [22]. This category includes a number of phrases that have multiple functions, but whose hedging function seems paramount (e.g. *there may be*, *to some extent*, *you might want to*). All of these formulas express some degree of qualification, mitigation, or tentativeness (Hyland 1998).

Other examples of hedges show clearly the tendency of these formulas to co-occur with other hedge words or phrases, as in B.1, where the formula is preceded by 'I mean, uh, you know'.

(B.1) but the, there are the examples of, and and the examples in the Renaissance I mean, uh, you know Copernicus is *to some extent* a figure of the Renaissance.

(2) Epistemic stance [32]. Epistemic stance formulas have to do with knowledge claims or demonstrations, expressions of certainty or uncertainty, beliefs, thoughts, or reports of claims by others.

(B.2) so we're just gonna be saying let's *assume that the* two variabilities in the two populations are the same . . .

(3) Obligation and directive [23]. Obligation and directive formulas are generally verb phrases directing readers or listeners to do or not do something, or to recall or attend to some observation, fact, or conclusion.

(B.3) Why? *Tell me what* your thought process is.

(4) Ability and possibility [29]. The ability and possibility formulas frame or introduce some possible or actual action or proposition. In the spoken genres, these formulas are often interactive phrases with the second person pronoun, as in *you can see*, *you can look at*, and *you're trying to*.

(B.4) We aren't *gonna be able to* predict all behaviors because chance variables play a big role.

(5) Evaluation [13]. The subcategory of evaluation is another addition to the Biber *et al.* taxonomy. Biber *et al*. included only two of these phrases and listed them under the category of impersonal obligation/directive (i.e. *it is important to*, *it is necessary to*). The AFL, however, includes several phrases that are clearly evaluative, without necessarily being directive, such as *the importance of*, *is consistent with*, *it is obvious that*, *it doesn't matter*. Furthermore, even those that are also directive we maintain function primarily as evaluators. Interestingly, of the thirteen phrases in this category, most are on the Written AFL; only one appears on the Core AFL (*the importance of*), and one on the Spoken AFL (*it doesn't matter*).

(B.5) Much macrosociological theory emphasizes *the importance of* societal variation.

(6) Intention/volition [11]. Most of the phrases in this category occur in the spoken genres, and express either the speaker's intention to do something, or the speaker's questioning of the listener's intention.

(B.6) So *let me just* take this off momentarily and put my other chart back on.

Group C: Discourse organizing expressions

Discourse organizers in the AFL fall into four main subcategories: metadiscourse, topic introduction, topic elaboration, and discourse markers. Each of these functions involves either signaling or referring to prior or upcoming

discourse. With the exception of the cause–effect subcategory of topic elaboration, all the discourse organizing expressions are more frequent in the spoken genres. This is consistent with Biber's (2006) finding that discourse markers are rare in written compared with spoken academic genres.

(1) Metadiscourse and textual reference [31]. The subcategory of discourse organizers with the largest number of phrases is the metadiscourse and textual reference category. As mentioned earlier, this functional category was not included in the Biber *et al*. taxonomy; most of the phrases we classified in this category were grouped in their study with the topic introduction/focus category (2004: 386). With no phrases on the Core AFL, these phrases are clearly differentiated between the spoken and written lists, thus indicating that metadiscourse formulas tend to be genre-specific.

(C.1) The seven studies are summarized *in the next section*.
(C.2) Yeah *I was gonna say* something similar to that.

(2) Topic introduction and focus [23]. This category overlaps functionally to a certain degree with the referring expressions identification and focus category. The main difference is that the global discourse organizing function of introducing a topic is primary here, with the phrase often framing an entire clause or upcoming segment of discourse, while the local referential function of identification is more salient for the other category.

(C.3) so the first thing we wanted to do was take *a look at* and see if in fact this compound can kill cancer cells.

(3) Topic elaboration. The topic elaboration subcategory includes two groups: non-causal topic elaboration, and cause and effect elaboration. Both categories function to signal further explication of a previously introduced topic.

(a) Non-causal [15]. Non-causal topic elaboration includes any phrase that is used to mark elaboration without any explicit causal relationship implied. This includes phrases that summarize or rephrase, as in *it turns out that* and *what happens is*, as well as interactive formulas and questions such as *see what I'm saying*, and *any questions about.*

(C.4) and let's just look at birth rate, and *what happens is* we have inverse, density dependence . . .

(b) Cause and effect [22]. The cause and effect formulas signal a reason, effect, or causal relationship. Although these are grouped as a subset of the topic elaboration formulas, they are an important functional group in and of themselves in academic discourse and for EAP teaching.

(C.5) at this point *in order to* get fired you have to do something really awful.

(C.6) *As a result*, research on the imposition of the death penalty in the United States has a long and distinguished history.

(4) Discourse markers [14]. The discourse markers category includes two subtypes. Connectives, such as *as well as*, *at the same time*, *in other words*, which connect and signal transitions between clauses or constituents. Interactive devices and formulas include *thank you very much*, *yes yes yes*, and *no no no*, which are phrases that stand alone and function as responses expressing agreement, disagreement, thanks, or surprise.

(C.7) Material data *as well as* functional principles must be taken into account for the physical design.

DISCUSSION AND CONCLUSIONS

Our methods and results suggest that formulaic sequences can be statistically defined and extracted from corpora of academic usage in order to identify those that have both high currency and functional utility. First, as in prior research with lexis (Nation 2001) and lexical bundles (Biber *et al*. 2004; Biber 2006), we used frequency of occurrence to identify constructions that appear above a baseline threshold frequency and which therefore have a reasonable currency in the language as a whole. Second, as in prior research defining academic lexis (Coxhead 2000), we identified those that appear more frequently in academic genres and registers and across a range of disciplines as being particular to EAP.

But currency alone does not ensure functional utility. However frequent in our coinage, nickels and dimes aren't worth as much as dollar bills. So too with formulas. When we assessed the educational and psycholinguistic validity of the items so selected, we found that they vary in worth as judged by experienced instructors, and in their processability by native speakers. In the present article, we show that experienced EAP and ESL instructors judge multiword sequences to be more formulaic, to have more clearly defined functions, and to be more worthy of instruction if they measure higher on the two statistical metrics of frequency and MI, with MI being the major determinant. In our companion paper (Ellis *et al*. 2008) we report experiments which showed how processing of these formulas varies in native speakers and in advanced second language learners of English.

Next, therefore, we used these findings to prioritize the formulas in our AFL for inclusion in EAP instruction using an empirically derived measure of utility that is both educationally valid and operationalizable with corpus linguistic metrics. Our FTW score weighs MI and frequency in the same way that EAP instructors did when judging a sample of these items for teaching worth. When we rank ordered the formulas according to this metric, the items which rose to the top did indeed appear to be more formulaic, coherent, and perceptually salient than those ordered by mere frequency or MI alone, thus providing

intuitive confirmation of the value of the FTW score. We used this ordering to inform the selection and prioritization for inclusion in EAP instruction of the Core and the top 200 Written and Spoken AFL formulas. This inclusion of MI for prioritizing such multiword formulas represents an important advance over previous research.

We then analyzed these formulas for discourse function to show that many of them fall into coherent discourse-pragmatic categories with enough face validity to encourage their integration into EAP instruction when discussing such functions as framing, identification and focus, contrast and comparison, evaluation, hedging, epistemic stance, discourse organization, and the like. Our AFL is categorized in this way in Table 4, with the functions further explained and exemplified in our Results section. It is our hope that this functional categorization, along with the FTW rank-ordered lists, will facilitate the inclusion of AFL formulas into EAP curricula, and that further work on the pedagogical value of the AFL will take these results as a starting point.

We recognize that there are other possible ways of going about this task, each with particular advantages and disadvantages. Biber *et al*.'s groundbreaking work in defining lexical bundles on the basis of frequency alone has served as a contrast for us throughout this paper. It showed how corpus analysis could be used to identify interesting EAP constructions. But it also showed how frequency alone generates too many items of undifferentiated value. Biber *et al*. (2004) included only four-word bundles because the same frequency cutoff would generate far too many lexical bundles to deal with if three- and five-word bundles were included; yet, as we show here, many of the important (and high FTW) words on our AFL are actually tri-grams. So too, many of the phrases in their high-frequency lexical bundles list don't appear in the AFL because while they gathered all strings of frequency in university teaching and textbooks, we used comparison non-academic corpora and the LL statistic to pull out only those phrases that are particularly frequent in academic discourse.

Our conclusions also stand in contrast to those of Hyland (2008) who argues that there are not enough lexical bundles common to multiple disciplines to constitute a core academic phrasal lexicon, and therefore advocates a strictly discipline-specific pedagogical approach to lexical bundles. Although we would not deny that disciplinary variation is important and worthy of further analysis, by using the metrics we did, we were able to derive a common core of academic formulas that do transcend disciplinary boundaries. Several factors that explain our divergent claims warrant mentioning. First, Hyland also analyzed only four-word bundles, whereas a glance at the top 50 Core AFL phrases shows the majority to be three-word phrases (e.g. *in terms of*, *in order to*, *in other words*, *whether or not*, *as a result*). Second, he used a higher cutoff threshold, whereas we started with a lower cutoff frequency; since our FTW score incorporates another statistic (MI) to insure relevance, the lower frequency range allowed us to cast a wider net without prioritizing numerous less relevant

formulas. Our research thus finds quite a number of core formulas common to all academic disciplines.

In closing, we are left with important conclusions relating to the complementarity of corpus, theoretical, and applied linguistics. Whatever the extraction method, there are so many constructions that there is ever a need for prioritization and organization. The current research persuades us that we will never be able to do without linguistic insights, both intuitive and academic. While some of these can be computationally approximated, as in the use of range of coverage of registers, and statistics such as MI and frequency in our FTW metric here, functional linguistic classification and the organization of constructions according to academic needs and purposes is essential in turning a list into something that might usefully inform curriculum or language testing materials.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistic* sonline.

NOTES

1 MICASE speech events include lectures, seminars, student presentations, office hours, and study groups; for further details about the specific genres in MICASE, see Simpson-Vlach and Leicher (2006). BNC spoken academic files include primarily lectures and tutorials. BNC written academic texts include research articles and textbooks.

2 Furthermore, this was the only corpus of conversational American English speech available to us; although telephone conversations are not necessarily ideal, they were quite adequate for comparison purposes.

3 Because these formulas appeared frequently in *both* spoken and written genres, the minimum threshold was set at six out of nine of the disciplinary sub-corpora, which had to include both written and spoken corpora. In fact, over 100 of the Core AFL formulas appeared in at least eight out of nine, and furthermore most of them occurred at frequencies well over 20 times per million.

REFERENCES

Barlow, M. 2004. *Collocate*. Athelstan Publications.

Barlow M. and **S. Kemmer** (eds). 2000. *Usage-Based Models of Language*. CSLI Publications.

Biber, D. 2006. *University Language*. John Benjamins.

Biber, D. and **F. Barbieri.** 2006. 'Lexical bundles in university spoken and written registers,' *English for Specific Purposes* 26: 263–86.

Biber, D., S. Conrad, and **V. Cortes.** 2004. 'If you look at . . .'': Lexical bundles in university teaching and textbooks,' *Applied Linguistics* 25: 371–405.

Biber, D., S. Conrad, and **R. Reppen.** 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and **E. Finegan.** 1999. *Longman Grammar of Spoken and Written English*. Pearson Education.

Bod R., J. Hay, and **S. Jannedy** (eds). 2003. *Probabilistic Linguistics*. MIT Press.

Brazil, D. 1995. *A Grammar of Speech*. Oxford University Press.

Bresnan, J. 1999. 'Linguistic theory at the turn of the century'. *Plenary address to the 12th World Congress of Applied Linguistics*. Tokyo, Japan.

British National Corpus. 2006. Available from http://www.natcorp.ox.ac.uk/. Accessed 1 October 2007.

Bybee, J. 2003. 'Sequentiality as the basis of constituent structure' in T. Givón and B. F. Malle (eds): *The Evolution of Language out of Pre-language*. John Benjamins.

Bybee J. and **P. Hopper** (eds). 2001. *Frequency and the Emergence of Linguistic Structure*. Benjamins.

Coxhead, A. 2000. 'A new Academic Word List,' *TESOL Quarterly* 34: 213–38.

Croft, W. and **A. Cruise.** 2004. *Cognitive Linguistics*. Cambridge University Press.

Ellis, N. C. 1996. 'Sequencing in SLA: Phonological memory, chunking, and points of order,' *Studies in Second Language Acquisition* 18/1: 91–126.

Ellis, N. C. 2002a. 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition,' *Studies in Second Language Acquisition* 24/2: 143–88.

Ellis, N. C. 2002b. 'Reflections on frequency effects in language processing,' *Studies in Second Language Acquisition* 24/2: 297–339.

Ellis, N. C. 2009. 'Optimizing the input: Frequency and sampling in usage-based and form-focussed learning' in M. H. Long and C. Doughty (eds): *Handbook of Second and Foreign Language Teaching*. Blackwell, pp. 139–58.

Ellis, N. C., E. Frey, and **I. Jalkanen.** 2009. 'The psycholinguistic reality of collocation and semantic prosody (1): Lexical access' in U. Römer and R. Schulze (eds): *Exploring the Lexis-Grammar Interface*. John Benjamins, pp. 89–114.

Ellis, N., R. Simpson-Vlach, and **C. Maynard.** 2008. 'Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL,' *TESOL Quarterly* 42/3: 375–96.

Flowerdew J. and **M. Peacock** (eds). 2001. *Research Perspectives on English for Academic Purposes*. Cambridge University Press.

Goldberg, A. E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Hunston, S. and **G. Francis.** 1996. *Pattern Grammar: A Corpus Driven Approach to the Lexical Grammar of English*. Benjamins.

Hyland, K. 1998. *Hedging in Scientific Research Articles*. John Benjamins.

Hyland, K. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. University of Michigan Press.

Hyland, K. 2008. 'As can be seen: Lexical bundles and disciplinary variation,' *English for Specific Purposes* 27: 4–21.

ICAME. 2006. Available from http://icame.uib.no/. Accessed 1 October 2007.

Jurafsky, D. 2002. 'Probabilistic modeling in psycholinguistics: Linguistic comprehension and production' in R. Bod, J. Hay, and S. Jannedy (eds): *Probabilistic Linguistics*. MIT Press.

Jurafsky, D. and **J. H. Martin.** 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Jurafsky, D., A. Bell, M. Gregory, and **W. D. Raymond.** 2001. 'Probabilistic relations between words: Evidence from reduction in lexical production' in J. Bybee and P. Hopper (eds): *Frequency and the Emergence of Linguistic Structure*. Benjamins.

Kuiper, K. 1996. *Smooth Talkers: The Linguistic Performance of Auctioneers and Sportscasters*. Erlbaum.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford University Press.

Lee, D. 2001. 'Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle,' *Language Learning & Technology* 5/3: 37–72.

Leech, L. 2000. 'Grammars of spoken English: New outcomes of corpus-oriented research,' *Language Learning* 50: 675–724.

Lewis, M. 1993. *The Lexical Approach: The State of ELT and The Way Forward*. Language Teaching Publications.

Manning, C. D. and **H. Schuetze.** 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

McEnery, T. and **A. Wilson.** 1996. *Corpus Linguistics*. Edinburgh University Press.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nattinger, J. R. and **J. DeCarrico.** 1992. *Lexical Phrases and Language Teaching*. Oxford University Press.

Oakes, M. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Pawley, A. and **F. H. Syder.** 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. C. Richards and R. W. Schmidt (eds): *Language and Communication*. Longman.

Rayson, P. and **R. Garside.** 2000. 'Comparing corpora using frequency profiling.' *Proceedings of the workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. Hong Kong.

Robinson, P. and **N. C. Ellis** (eds). 2008. *A Handbook of Cognitive Linguistics and Second Language Acquisition*. Routledge.

Schmitt, N. (ed.). 2004. *Formulaic Sequences*. Benjamins.

Simpson, R. 2004. 'Stylistic features of academic speech: The role of formulaic expressions' in T. Upton and U. Connor (eds): *Discourse in the Professions: Perspectives from Corpus Linguistics*. John Benjamins.

Simpson, R. and **D. Mendis.** 2003. 'A corpus-based study of idioms in academic speech,' *TESOL Quarterly* 3: 419–41.

Simpson, R., S. Briggs, J. Ovens, and **J. M. Swales.** 2002. *The Michigan Corpus of Academic Spoken English*. The Regents of the University of Michigan.

Simpson-Vlach, R. and **S. Leicher.** 2006. *The MICASE Handbook: A resource for users of the Michigan Corpus of Academic Spoken English*. University of Michigan Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. Routledge.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Switchboard. 2006, August 5, 2006. '*A user's manual*.' Available from http://www.ldc.upenn.edu/Catalog/docs/switchboard/.

Tomasello M. (ed.). 1998. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Erlbaum.

Tomasello, M. 2003. *Constructing a Language*. Harvard University Press.

West, M. 1953. *A General Service List of English Words*. Longman.

Wray, A. 1999. 'Formulaic sequences in learners and native speakers,' *Language Teaching* 32: 213–31.

Wray, A. 2000. 'Formulaic sequences in second language teaching: Principle and practice,' *Applied Linguistics* 21: 463–89.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. and **M. R. Perkins.** 2000. 'The functions of formulaic language: An integrated model,' *Language and Communication* 20: 1–28.